The Primary Review

... children, their world, their education

INTERIM REPORTS

Research Survey 3/4

THE QUALITY OF LEARNING: ASSESSMENT ALTERNATIVES FOR PRIMARY EDUCATION

Wynne Harlen
University of Bristol

For other interim reports in this series, and for briefings on each report, go to www.primaryreview.org.uk

This report has been commissioned as evidence to the Primary Review. The analysis and opinions it contains are the authors' own.

Esmée Fairbairn FOUNDATION

UNIVERSITY OF CAMBRIDGE
Faculty of Education

PRIMARY REVIEW INTERIM REPORTS

THE QUALITY OF LEARNING:
assessment alternatives for primary education

Primary Review Research Survey 3/4

Wynne Harlen

October 2007

This is one of a series of 32 interim reports from the Primary Review, an independent enquiry into the condition and future of primary education in England. The Review was launched in October 2006 and will publish its final report in late 2008.

The Primary Review, supported by Esmée Fairbairn Foundation, is based at the University of Cambridge Faculty of Education and directed by Robin Alexander.

A briefing which summarises key issues from this report has also been published. The report and briefing are available electronically at the Primary Review website: www.primaryreview.org.uk. The website also contains Information about other reports in this series and about the Primary Review as a whole. (Note that minor amendments may be made to the electronic version of reports after the hard copies have been printed).

We want this report to contribute to the debate about English primary education, so we would welcome readers' comments on anything it contains. Please write to: evidence@primaryreview.org.uk.

The report forms part of the Review's research survey strand, which consists of thirty specially-commissioned surveys of published research and other evidence relating to the Review's ten themes. The themes and reports are listed in Appendices 1 and 3.

This survey relates to Primary Review theme 3, **Curriculum and Assessment**.

The author: Professor Wynne Harlen is Visiting Professor of Education at the University of Bristol and former Director of the Scottish Council for Research in Education.

Suggested citation: Harlen, W. (2007) *The Quality of Learning: assessment alternatives for primary education.* (Primary Review Research Survey 3/4), Cambridge: University of Cambridge Faculty of Education.

Published October 2007 by The Primary Review,
University of Cambridge Faculty of Education,
184 Hills Road, Cambridge, CB2 8PQ, UK.

British Library Cataloguing in Publication Data:
A catalogue record for this publication is available from the British Library.

ISBN 978-1-906478-03-2

THE QUALITY OF LEARNING:
ASSESSMENT ALTERNATIVES FOR PRIMARY EDUCATION

Introduction

Why and how we assess our pupils has an enormous impact on their educational experience and consequently on how and what they learn. This paper provides a critical review of the assessment system in England in the light of evidence from research and practice. It begins by considering the 'why' and 'how' of assessment and then, in section 3, describes how the various purposes and uses of assessment are met in England, in the other countries of the UK and in France, Sweden and New Zealand. In the fourth section alternative methods of conducting pupil assessment for different purposes are considered in relation to their validity, reliability, impact on learning and teaching and cost. The main points from this analysis are drawn together in the final section, indicating viable alternatives to tests and to the high stakes use of measures of pupil achievement.

Note on terminology

At the start it is perhaps necessary to make clear that the word 'assessment' is used here to refer to the process of making judgements about pupils' learning - and more generally about any learner's learning. In some countries, including the USA, the word 'evaluation' is used for this process and in many cases the two words are used interchangeably. Here we use the word evaluation to refer to the process of making judgements about teaching, programmes, systems, materials, and so on. Both assessment and evaluation involve decisions about what evidence to use, the collection of that evidence in a systematic and planned way, the interpretation of the evidence to produce a judgement, and the communication and use of the judgement; it is the type of evidence that defines the difference. The evidence, of whatever kind, is only ever an indication, or sample, of a wider range that could be used.

1. Why assess?

There are two main reasons for assessing pupils:

- to help their learning
- to report on what has been learned.

These are usually discussed as different purposes of assessment and sometimes, mistakenly, as different *kinds* of assessment and ones that are somehow opposed to one another. They are certainly different in several important respects, but what should unite them is the aim of making a positive contribution to learning. This impact on learning is one of the criteria to be used later in evaluating different answers to the question of how we assess.

Decisions that are involved in assessment, about the evidence to gather, how it is judged and by whom, how the results are used and by whom, follow from the reasons for the assessment. Assessment for first of the two reasons above is called formative assessment or alternatively, assessment *for* learning. It is defined as:

> the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.
>
> Assessment Reform Group (ARG) (2002a)

It is carried out as part of teaching and so involves the collection and use of evidence about the learning in relation to the specific activities and goals of a lesson. This is detailed evidence, interpreted by the teacher and pupil to decide where the pupil has reached and so what next steps are needed to help achievement of the goals, or to move on.

Assessment for the second reason is called summative, or assessment *of* learning, and is carried out for the purpose of reporting achievement of individual pupils at a particular time. It relates to broad learning goals that are achieved over a period of time. It can be conducted in various ways, as discussed later, including by tests or examinations at a certain time, or summarising achievement across a period of time up to the reporting date.

Uses of assessment results

Before going on to the question of 'how', it is necessary to consider the use made of the results since this influences decisions about how to gather and interpret evidence. For formative assessment there is, by definition, one main use of the data, to help learning. If the information about pupils' learning is not used to help that learning, then the process cannot be described as formative assessment. By contrast, the data from summative assessment are used in several ways, some relating to individual pupils and some to aggregated results of groups of pupils.

For individual pupils, the uses of summative assessment can be described as either 'internal' or 'external' to the school community:

- 'Internal' uses include using regular grading, record keeping and reporting to parents and to the pupils themselves; at secondary level, informing decisions about courses to follow where there are options within the school.
- 'External' uses include meeting the requirements of statutory national assessment, for selection, where selective secondary schools exist; at the secondary level, certification by examination bodies or for vocational qualifications, selection for further or higher education.

In addition to these uses, which relate to making judgements about individual pupils, results aggregated for groups of pupils are used for evaluating the effectiveness of the education provided by teachers, schools and local authorities. The main uses of aggregated results in England are:

- Accountability: for evaluation of teachers, schools, local authorities;
- Monitoring: to compare results for pupils of certain ages and stages, year on year, to identify change in 'standards'.

Assessment systems

The use of individual pupil results for accountability and monitoring is strongly contested and is a matter to which we return later. However, at this point it is useful to note that any *system of assessment* has to identify the role that measures of pupil performance will take in the accountability of teachers and schools and in monitoring at local and national levels, as well as how evidence about individual pupils will be gathered and used for different purposes.

As in any system, these various parts are interconnected and how one part is carried out influences how other parts can function. A prime example of this interaction is seen when schools are held accountable for meeting targets set solely on the basis of the results of pupils' performance in external tests. There is evidence at the primary level from the PACE

(Primary Assessment Curriculum and Experience) project that this is associated with teachers' own classroom own assessment becoming focused on achievement rather than learning (Pollard et al 2000; Pollard and Triggs 2000). Other interactions among elements within assessment systems become clear when we consider different systems in section 3.

The 'stakes' of assessment

The term 'high stakes' has been adopted to refer to pupil assessment where the results are used to make important decisions, either for the pupil or for the teacher, or both. In the case of primary school pupils, the stakes are high in places where there are selective secondary schools and entrance to a preferred school depends on the outcomes of assessment. Even where the 11+ examination has been ended, as is now the case in Northern Ireland, the assessment that replaces it takes on the high stakes. However, a far more widespread source of high stakes is the use of the results of national tests for the evaluation of schools. Although the results of national tests may not, in theory, have high stakes for pupils, the results are of considerable importance for teachers where, as in England, aggregated results are used to set targets which schools are held accountable for meeting. The consequences of pupils not achieving at certain levels can be severe, including the school being described as having 'serious weaknesses', being placed in 'special measures' or even closed. To avoid these consequences, inevitably teachers place emphasis on making sure that pupils' test results are maximised, with all that this implies for teaching to the test and giving practice tests (ARG 2002b).

As discussed further in section 4, the optimistic view that a range of purposes can be served by using the data from a single source is the root cause of the negative impact of testing on the curriculum and pedagogy. The use of national test results for individual school accountability, for monitoring national standards and for reporting on individual pupils means that the information is not well matched to what is required for each of these purposes. Although this was, indeed, what the Task Group on Assessment and Testing (TGAT) report (Department of Education and Science/ Welsh Office 1988) suggested, it was based on expectations that Black, the Task Group chair, later described as naïve: that 'the assessment results (would) be interpreted in a context of interpretation so that they would not mislead those they were meant to inform' (Black 1997: 41).

2. How should we assess?

Decisions about how the evidence for assessment is gathered, about the basis for judgement, and about what quality assurance procedures need to be in place are made in the light of how the results are to be used. Before looking at the criteria for evaluating different ways of assessing pupils, some options in making these decisions are briefly considered.

What evidence?

In theory anything that a pupil does provides evidence of some ability or attribute that is required in doing it. So the regular work that pupils do in school is a rich source of evidence about the abilities and attributes that the school aims to help pupils develop. This evidence is, however, unstructured and varies in some degree from class to class, even pupil to pupil. These differences can lead to unfairness in the judgements unless the assessment procedures ensure that the judgements of equivalent work are comparable. One way to avoid this problem entirely is to create the same conditions and tasks for all pupils; that is, to use tests.

Testing is a method of assessment in which procedures, such as the task to be undertaken and often the conditions and timing, are specified. Usually tests are marked using a scheme prescribed either by the pupils' teacher or external markers, who are often teachers from

other schools. The reason for the uniform procedures is to allow comparability between the results of pupils, who may take the tests in different places. Tests are described as 'performance', 'practical', 'paper-and pencil', 'multiple choice', 'open book', etc. according to the nature of the tasks that are prescribed.

Teachers regularly create their own tests for internal school use; in other cases they are created by an agency external to the school. Tests are criticised on a number of points, considered in more detail later, but it is the emotional reaction of many pupils to them that is a considerable cause of concern. The specific tasks or items are unknown beforehand and pupils have to work under the pressure of the allowed time. This increases the fear that they will 'forget everything' when faced with the test; the anticipation is often as unpleasant as the test itself. To counter this, and also to assess domains that are not adequately assessed in written, timed tests or examinations, assessment tasks may be embedded in normal work. The intention is that these tasks are treated as normal work. It may work well where the use is internal to the school, but the expectation of 'normality' is defeated when the results are used for making important decisions and the tasks become the focus of special attention by teacher and pupils.

How is evidence turned into a judgement?

Making a judgement in assessment is a process in which evidence is compared with some standard. The standard might be what other pupils (of the same age or experience) can do. This is *norm-referencing* and the judgement will depend on what others do as well as what the individual being assessed does. In *criterion-referencing* the standard is a description of certain kinds of performance and the judgement does not depend on what others do, but only on how the individual's performance matches up to the criterion. In *pupil-referenced*, or ipsative, assessment the pupil's previous performance is taken into account and the judgement reflects progress as well as the point reached. The judgements made of different pupils' achievements are then based on different standards, which is appropriate when the purpose is to help learning but not for summative purposes.

Summative assessment is either criterion-referenced or norm-referenced. Criterion-referencing is intended to indicate what pupils, described as having reached a certain level, can do. Formative assessment is often a mixture of ipsative and criterion referencing, where pupils are given feedback that takes into account the effort they have put in and the progress made as well as what has been achieved.

Who makes the judgement?

It is the essence of formative assessment that the information is collected and used in relation to on-going activities. Thus it is the teacher, together with the pupil, who collects and judges the information about what is being learned. In some cases it may be possible for teacher and pupil together to decide on immediate action. In other cases, the teacher may take note of what is needed and provide for it at a later time.

In summative assessment where external tests are used the judgements will be made by someone outside the school, usually a teacher who has been trained to apply a mark scheme or to use level descriptions (criteria), to decide the 'level' that can be awarded. Teachers can also take a more central role in the assessment of their own pupils by collecting evidence and making judgements about the levels achieved. Judging a range of work against criteria is not a straight-forward matter of relating evidence to description (Wilmut 2004). For example, the level descriptions of the national curriculum assessment comprise a series of general statements that can be applied to a range of content and contexts in a subject area. Not all criteria will apply to work conducted over a particular period, and there will be

inconsistencies in pupils' performance – meeting some criteria at one level but not those at a lower level, for instance. Typically the process of using criteria involves going to and fro between the statements and the evidence, and some trade off between criteria at different levels; all of which involve some value judgements. Quality assurance procedures come into play to minimise the differences between teachers' judgements of the same work.

How is quality assured?

Quality assurance, meaning procedures to minimise inaccuracy due to any of a range of causes, has a role in parts of all types of summative assessment. In the case of external tests, some quality assurance is built into the test development process when items and procedures are modified as a result of trials. Where teachers make judgements of pupils' work at the primary school level, quality assurance may take the form of group moderation, or the use of examples of assessed work to guide decisions, or the use of items from a bank of tests and tasks that have been calibrated in terms of levels of achievement. The purpose is to align the judgements of different teachers. When the process involves teachers meeting to review samples of pupils' work it has value beyond the reliability of the results (ARG 2006a). The rigour of the moderation process that is necessary depends on the 'stakes' attached to the results. Where the stakes are relatively low, as in internal uses of summative assessment, within-school moderation meetings are adequate, whilst inter-school meetings are needed when the results are used for external purposes. The use of exemplification and items banks can be seen as substitute for moderation meetings, however, thus reducing opportunities for inter-school discussions and for the professional development that these meetings can have. (There is more discussion of these quality assurance alternatives in section 5.)

At the secondary level where certification depend on teachers' judgements, in part or whole, the range of procedures used includes visits to the school of verifiers and moderators, inspection of samples of work and statistical adjustment of marks (Harlen 1994). It is too early for the impact of recent suggestions for accrediting schools (ACCAC[1] 2004) or teachers, at the secondary level (ACSL 2006), or for the newly set up Institute of Educational Assessors (IEA 2006), to be considered.

3. Some examples of assessment systems

In this section we describe briefly some key aspects of the assessment system at the primary level in England and, for comparison, in the other countries of the UK and in New Zealand, Sweden and France. In each case we consider how the system provides for formative and summative use of individual pupil assessment, for school evaluation and for the national monitoring of standards.

England

Assessment begins in the 'foundation stage', the period when children may be in nursery education or in the reception year of a primary school. The foundation stage ends when children enter Year 1 of primary education in the September following their fifth birthday. In order to provide 'a way of summarizing young children's achievements at the end of the foundation stage' (QCA 2003), the Foundation Stage Profile (FSP) was introduced in the school year 2002/3. The FSP comprises 13 scales relating to personal, social and emotional development, communication, language and literacy, mathematical development, knowledge and understanding of the world, physical development and creative development. For each scale a judgement is made in terms of nine points, relating to the

[1] Formerly the Qualifications, Curriculum and Assessment Authority for Wales, now within the Department for Children, Education, Lifelong Learning and Skills (DCELLS) of the Welsh Assembly Government.

child's progress towards achieving the 'early learning' goals. It is intended that the profile is built up over the foundation stage so that the evidence can be used formatively and then summarised against the performance descriptions of the scales for reporting at the end of each term. The process is entirely teacher-based and the evidence for completing the profile is derived from on-going learning activities. Occasionally, additional observations (of behaviour in different contexts) may be required although these should still be situated within the normal curriculum provision.

At present the FSP assessments cannot be used to make comparisons between schools in the same way as national test and examination results used in England, since only aggregated results are submitted to the DfES by local authorities and results for specific schools cannot be identified. Nevertheless, local authorities are still able to produce comparative information for schools and the results from individual schools or settings can be compared with national data at the time of inspections. There are also indications that DCSF/DfES policy on collecting individual pupils and school data may change in 2007 thus opening the possibility of results being used for accountability.

The teacher-based, on-going, wide-ranging, low stakes assessment of the FSP contrasts in many ways with what pupils experience in the primary school in England. At the end of Key Stage 1 (years 1 and 2, pupils aged 5-7) and of Key Stage 2 (years 3 to 6, pupils aged 7-11) there are external tests and tasks in English and mathematics (and in science at Key Stage 2 only) that teachers are required to administer in a strictly controlled manner. In addition to the core subject tests at the end of Key Stages 1 and 2, assessment by teachers is also required. For Key Stage 2 both test results and teachers' assessment results are reported and are said to have equal status. From 2005, at Key Stage 1 only the teachers' assessment results are reported but tests in English and mathematics still have to be given to inform the teachers' judgements.

Although it is only at the end of a Key Stage that pupils' performance must be reported in terms of national curriculum levels, schools have a statutory requirement to provide a summative report for parents for each pupil and each subject studied at least once every year and schools often choose to include the levels judged to have been reached. This trend towards annual reporting in terms of levels has been reinforced by widespread use of the optional tests produced by QCA for years between the end of Key Stages for the core subjects.

The frequency of testing is set to increase further following the proposal of single-level tests in the consultation document entitled *Making Good Progress* (DfES 2007). This proposes the introduction of new tests, for pupils in Key Stages 2 and 3, designed to assess achievement at a particular level. These tests would be shorter than the current end of Key Stage tests and in mathematics and English only. Pupils would sit a test when their teacher judged them to be able to pass. Testing opportunities would be given twice a year, in December and June, beginning in December 2007. It is proposed that the results of the tests would be the basis of 'progression targets' for teachers and schools, adding to the targets based on end of Key Stage tests. The progress measure would be 'the percentage of pupils who make two levels of NC progress during Key Stage two'. Thus is it clear that these proposed would be used in the evaluation of teachers and schools, adding considerably to the pressures felt by teachers and pupils. There is further discussion of these proposals and evidence of the impact of testing in Section 4 .

The formative use of assessment at the primary level features prominently in the Primary Strategy, where assessment for learning forms part of the new primary resource *Excellence and Enjoyment: Learning and teaching in the primary ye*ars (DfES 2004). The renewed Primary Framework for Literacy and Mathematics also urges better use of assessment. However,

implementation of assessment for learning, which is voluntary, is limited by the attention that teachers feel needs to be given to ensuring that the statutory test results are optimised, since it is only national test results that are used to create targets for schools and give rise to league tables. The results from the same end of Key Stage tests are used to evaluate the performance of schools, local authorities and to monitor changes in the performance of pupils year on year in the country as a whole.

Scotland, Wales and Northern Ireland

In these three countries of the UK considerable changes are underway, or being considered, in the systems of assessment created in the early 1990s. Scotland, having begun major reforms with a review of assessment in 1999, has gone furthest in implementing change. Wales is in the process of phasing in change and, as of 2006, in Northern Ireland sweeping policy and organisational changes are in train. However, while these countries are at different points in implementation of change and differ in the details of the change, there is sufficient in common in the direction of the changes, towards greater use of assessment by teachers and away from frequent testing, to warrant discussing them together.

Scotland

Scotland is the largest of these three countries, with about 2,200 primary, 385 non-selective secondary, 57 independent secondary schools and 190 special schools. Transfer from primary to secondary school takes place at the end of year 7 (P7), so there are seven years of primary education and four of secondary education before the statutory school leaving age of 16. Neither the curriculum nor its assessment is governed by legislation in Scotland, as it is in the rest of the UK. In the absence of regulation, factors which ensure implementation of changes include a tradition of conforming with central policy and wide consultation on changes. Inevitably, achieving consensus is a slow process and often means that change is evolutionary.

The newly introduced system of assessment in primary schools in Scotland contrasts sharply with that across the border in England. This has come about in reaction to the practice that developed through the 1990s after the introduction of the national assessment. Despite the initial intention in the assessment guidelines introduced by the Scottish Education Department in 1991 (SED 1991) to give a strong role to teachers' professional judgement and the formative use of assessment, there was, as in other countries of the UK, an increasing emphasis on standards, target-setting and accountability in the mid- to late-1990s that distorted the curriculum and moved the focus of assessment to measurement (Hutchinson and Hayward 2005). HMI reports showed that the intention that national tests should be used to moderate teachers' professional judgements was not being realised. Instead targets were dominating classroom assessment practice and tests were used to decide the level of pupils' achievements.

In response, the Minister for Education in the newly formed Scottish Parliament commissioned a national survey on Assessment 3-14.. The report, arising from the analysis of responses from a wide group of stakeholders, identified several major areas for change (Hayward, Kane and Cogan 2000). As a result a major programme of reform in assessment, entitled '*Assessment is for Learning*', was introduced in 2003. The programme was concerned with the whole system of assessment for the age range 3 to 14. It was recognised that major changes would only be possible if policy-makers, researchers and teachers worked together to collectively own the new procedures. Thus new procedures to promote and sustain change were developed collaboratively, with groups of schools working together. Ten projects were set up to this end, between them dealing with formative assessment, personal

learning plans for pupils, moderation of teachers' assessment, the development of a bank of tests and tasks for moderation of teachers' judgements and a framework for reporting progress to parents and others. Almost all local authorities (30 out of 32) took part in the development of at least one project, and by the end of 2004 over 1,500 schools were involved. On completion of the development programme, *Assessment is for Learning* was formally adopted as policy for the education of pupils aged 3-14 by ministers (SEED 2004) and the action proposed included ensuring the participation of all schools by 2007.

The main features of the programme are as follows:

- Formative assessment is in operation both for pupils and for staff, with particular emphasis on self-assessment, setting own goals and reflecting on learning.
- For summative assessment, teachers use a range of evidence from everyday activities to check on pupils' progress. There are no Key Stages in Scotland and pupils are assessed by their teachers as having reached a level of development (identified in the curriculum guidelines by criteria at six levels, A to F), using evidence from regular activities. Assessment against the level criteria is an on-going process; a pupil may be judged to have reached a level at any time. When confirmed by moderation, this is recorded and then reported at the appropriate time.
- Quality assurance of teachers' judgements of pupils' performance is through taking part in collaborative moderation within and across schools to share standards and/or using National Assessment. A circular (SEED, 2005a) advising on practical implications of the implementation of the programme described the use of tests as 'Another way for teachers to check their judgements against national standards'. Teachers can use an externally devised bank of assessments and tests and compare the results with the results of their own classroom assessments, when they judge that children have reached a particular level (SEED Circular 02, June, 2005a).
- For monitoring of national standards there is a separate rolling programme of assessment of a sample of pupils, now called the Scottish Survey of Achievement. Begun in 1983 as the Assessment of Achievement Programme, it was revised in 2003 to include four subjects; English, mathematics, science and social subjects, each assessed in turn once every four years. Samples of pupils in years P3, P5, P7 and S2 (8, 10, 12 and 14 years of age) are tested in each survey (SEED 2005b).
- For evaluation of schools, a school self-evaluation toolkit has been developed to support self-evaluation against quality indicators, which include, but are not confined to, pupil performance data (HMIe 2006).

Wales

Wales has about 1500 state primary schools for years 1 to 6, from which pupils transfer at the age of 11 to the 230 non-selective secondary schools (there are no selective secondary schools and no middle schools). The curriculum and assessment in place until 2000 were established by the same Education Reform Act of 1988 that applied to both England and Wales. In 2000, following several reviews of the curriculum, the Wales Curriculum 2000 was introduced and the decision was taken to end statutory tests and tasks at the end of Key Stage 1. From that date, statutory assessment by teachers was the only form of assessment at the end of Key Stage 1 and at the end of Key Stage 2 both teachers' assessment and results of tests, intended to be of equal status, were reported. Whilst it has not been the practice in Wales to publish performance tables based on test results for individual schools, the results of both teachers' assessment and national tests were published as summaries for each subject and for each LEA and for Wales as a whole. ACCAC (then the Qualifications, Curriculum and

Assessment Authority for Wales, now within the Department for Children, Education, Lifelong Learning and Skills (DCELLS) of the Welsh Assembly Government) also published guidance materials to improve the consistency of teachers' assessment.

A review of the school curriculum and assessment arrangements, begun in 2003, recommended more sweeping changes in the assessment system (ACCAC 2004). These were largely accepted by the Minister of Education in the Welsh Assembly Government. The main changes proposed were:

- Tests at the end of Key Stage 2 were to be phased out and from 2005 the assessment of levels reached by pupils was to be based only on 'best fit' judgements by teachers.
- A system of moderation was to be set up in order to ensure an acceptable level of consistency in teachers' judgements. Schools were to be grouped on the basis of notional secondary school catchment areas with each primary school linked for the purpose of these moderation procedures to a particular secondary school. Primary and secondary teachers from each group of schools would meet twice in each school year for agreement trials using pupils' work in the subjects being assessed.
- Tests at the end of Key Stage 3 were also to be discontinued and the reporting of end-of-Key Stage assessment in all subjects to be based on teachers' 'best fit' judgements in relation to national curriculum levels.
- The use of data about pupils' performance would be only one element used in school self-evaluation.
- The use of data about pupils' performance would be only one element in the monitoring of overall performance at local authority and national levels.

There has been considerable effort in supporting schools in setting up procedures to assure quality in teacher assessment outcomes. This has included centrally produced guidance of using professional judgements, which is intended to move teachers away from dependence on test-derived data. It is recognised that it will take time to build up trust in teachers' judgements and convince them that the different use of time is worthwhile. As in Scotland, the involvement and sense of ownership of new arrangements will be an important factor in helping teachers through the period of change.

Northern Ireland

Although a smaller country than Wales, Northern Ireland, like Scotland, has a long tradition of a separate education system. The body currently responsible for the curriculum and assessment has, since 1994, been the Council for Curriculum Examinations and Assessment (CCEA), a non-departmental body reporting to the Department of Education in Northern Ireland. However, an extensive reorganisation of the administration of education in Northern Ireland has been set in train, in which a single body (the Education and Skills Authority) will take over the functions of the existing Library Boards (local authorities), the CCEA and the Regional Training Unit. At the same time there are plans (published in the Bain report 2006) to close a number of schools (around one third of the current number) in order to remove places left empty in a declining population and in the process to give preference to religiously integrated schools.

The curriculum is described in terms of Key Stages, but these are different from the Key Stages in England. Children move from pre-school into year 1 in the year in which they reach the age of 5, not after it, so they are on average younger than year 1 children in the rest of the UK. Foundation stage refers to years 1 and 2, Key Stage 1 to years 3 and 4, and Key Stage 2 to years 5 to 7, with pupils moving into secondary school at the age of 11/12. Secondary

education is selective and the selection mechanism, the transfer test known as the 11+ examination, has been a defining feature of Northern Ireland education since 1947. The 11+ has not only dominated the curriculum in years 6 and 7 but has both sustained and been sustained by the prevailing 'testing culture'.

Several reviews of the curriculum and assessment (Harland et al 1999a, 1999b, 2001 and 2002) and of the selection system in particular (Johnston and McClune 2000; Leonard and Davey 2001; Gardner and Cowan 2005) between them highlighted a number of problems of the assessment system. These include the absence of quality assurance of teachers' judgements at the end of Key Stage 1, where there are no tests and where the reporting of performance is on the basis of assessment by teachers. In Key Stage 2 there are no national tests as such but teachers are required to use certain external Assessment Units provided by CCEA to moderate their assessment at the end of the Key Stage. However, instead of being used to confirm teachers' judgements, it has been found that these tasks are frequently administered as tests and used to determine the level at which children are working. Moderation of teachers' judgements by CCEA is not felt to be sufficiently rigorous and teachers do not trust the judgement of other teachers and schools, particularly where there is competition to attract pupils in a shrinking catchment area. Moreover, Key Stage 3 teachers put little faith in the assessment of the primary teachers. There is little use of assessment to help learning.

Recognition of these problems has led to recommendations for change including the ending of the 11+ transfer tests. Although this will not change the need for selection, it is expected that other major features in a revised system, such as assessment by teachers and the formative use of assessment, will mean that primary children and their parents are better prepared for making realistic decisions about the appropriate secondary school. In new arrangements being planned, all summative assessment at Key Stages 1, 2 and 3 will be teacher based and moderated on a three year cycle. Several approaches to quality assurance and quality control of teachers' judgements are being considered, including the accreditation of schools, moderation of procedures, and professional development in assessment techniques.

A Pupil Profile has been developed by CCEA in collaboration with parents, teachers and other educational partners. The aim of the Profile is to provide a record of each individual pupil's achievements, from on-going assessment, in a way that provides consistency between schools and reflects the revised Northern Ireland curriculum, which will become statutory in 2007. This means that achievements will be recorded for Communication (reading, speaking, listening, responding and presenting), Using mathematics, ICT, Thinking skills and personal capability (being creative, problem solving, self management, working with others) as well as for subject areas of the language, mathematics, the arts, the world around, and so on. When phased in from 2007, it should help to avoid the current narrow focus on reading, writing and mathematics.

Countries outside the UK

New Zealand

New Zealand, with a population of about four million, is quite similar to Scotland in having a high proportion of small primary schools. Most pupils attend primary school for six years, although there are some schools covering the first eight years of schooling. In 1991 the curriculum was restructured in some 'sweeping changes modelled on the curriculum and assessment changes in the late 1980s in England and Wales' (Crooks 2002: 239). Thus there are strands within subject areas and achievement objectives at eight levels within each

strand. The first five levels are spaced about two years apart. This curriculum is now being reviewed and is likely to be replaced by a less detailed description of objectives.

The attempts by teachers, encouraged by the Ministry of Education and the inspectorate (the Education Review Office), to use the levels of the curriculum in recording and reporting their assessment of pupils met with similar problems to some of those encountered in England. The levels are too widely spaced to give a satisfying account of progress, teachers' judgements of similar work varied and there was a tendency to base judgements on too narrow a range of tasks. Making the objectives more specific by sub-dividing them into different components or by creating intermediate steps within levels may improve agreement between teachers but may reduce the validity of the assessment. These difficulties remain but various forms of help with assessment have been developed for teachers. One of these provides assessment materials to be used with pupils on entry to school. These individually administered tasks are designed to give diagnostic information about certain aspects of numeracy, oracy and written language. A second support takes the form of banks of tasks, freely available on the Internet, for assessing English, mathematics and science at primary level. Teachers use these in various ways to check their judgement of levels, as ideas for their own assessment tasks and to check their pupils' achievement against national norms. There are also exemplars of performance at the different levels for each strand and subject area. Other free assessment materials for upper primary and lower secondary pupils are available from the Internet.

To date the introduction of national tests for all pupils has been resisted, although it has been proposed by the National Party (Government of New Zealand 1998). Instead, what pupils experience in terms of tests or tasks is a matter for their teacher and school. The assessment results are used within schools for monitoring progress and standards and by the inspectorate for external review, but these uses, according to Crooks, 'do not have a dominant influence on teachers' assessment practice' (Crooks 2002: 246). He concludes that the assessment in New Zealand primary schools

> is predominantly low stakes assessment focused on monitoring pupils' learning, improving learning through direct feedback to students or adjustments to teaching programmes. Written or oral reports to parents can be seen as complementing the formative role by giving guidance to parents and students, while also having a summative role.
>
> Crooks (2002) p 246

The low stakes is preserved by the existence, as in Scotland, of a quite separate programme for national monitoring. The National Education Monitoring Project (NEMP 2006), which has been in existence since 1993, assesses each year small samples of pupils using tasks that are administered by teachers seconded from other schools and trained to administer the tasks. Over a four year cycle, 15 different curriculum areas are assessed and reported. This wide range ensures that attention is not focused on areas of learning that are most easily assessed by tests.

Sweden

Sweden provides an interesting contrast with England in terms of trends over the last 20 years. While, in the late 1980s, England was moving from a decentralised to a centralised curriculum and assessment system, the reverse was the case in Sweden. Reforms, begun in 1980 and continued through the 1990s, gave more decision-making power and financial responsibility at the local level. The 290 municipalities now have freedom to decide the courses and curriculum they offer, while the central government provides guidelines and general regulations. Thus 'it is the responsibility of the municipalities and the school board of each municipality to formulate educational plans for their school district and ensure that

these plans are carried out in practice (National Agency for Education 2005)' (Wikstrom, 2006: 115). The National Agency for Education provided guidance to municipalities as to the implementation of government guidelines.

The population of about nine million in Sweden is relatively homogeneous and the schools and school outcomes are more even across the country than in many other OECD countries (Wikstrom 2006). School is optional at age six, but compulsory from age seven. For the first nine years, pupils are in comprehensive schools, when more than 90 per cent move to upper secondary school at the age of 16.

As well as school autonomy, assessment changed radically in the 1990s. Until the reform of 1994, pupils' grades were norm-referenced.

> Objectives to be taught were described in centrally issued curricula that contained rather detailed descriptions of what type of knowledge should be the focus of each subject. Students were then awarded a grade from 1 to 5, on a scale representing the overall achievement in the country.
>
> Wikstrom (2006) p 117

The usual problems of norm-referencing were experienced: teachers mistakenly assumed that each class should be graded according to a normal distribution; the meaning of each grade level could change from one year to another and so they could not be used for monitoring over time; grades gave no information about what pupils could do.

Although norm-referencing served quite well the purpose of selection at the end of upper secondary school, its use at earlier stages was out of line with the need for grades to be informative and was also in conflict with a society that values equal opportunities rather than competition for grades. From 1994, criterion-referencing was introduced in the comprehensive school. All assessment is carried out by teachers. In the early years grades are not assigned, although teachers report to parents at least twice a year. Only in the upper years of the comprehensive school are pupils graded according to how well their work meets the criteria at various levels for each subject. As in many statements of curriculum objectives, the description of the levels is in quite general terms and teachers are expected to collaborate in agreeing their operational meaning. Wikstrom comments that, 'since it is the teachers who assess and grade the students there has been no need for standardised examination tests, and the idea of using tests for such purposes has not been discussed for several decades' (Wikstrom 2006:120/1), a logic that has escaped politicians and some commentators alike in England!

Nevertheless there are tests for 'scale calibration' available in Swedish, English and mathematics which are described as National Tests. Teacher use these tests in the upper comprehensive school as part of the evidence on which they base their grades. The tests are also intended to have a diagnostic function. However, the lack of clear moderation procedures for aligning grade judgements by teachers means that the results are of low reliability and so cannot be used to monitor standards over time. Recognition of these problems has led to plans to improve teachers' assessment practice through more effective teacher education in assessment. It is the role of the National Agency for Education to provide the municipalities with guidance for aligning grades. It is also responsible for monitoring the system and the inspection of schools.

France

With a population of about 60 million, France has a school structure of primary schools up to the age of 10/11, lower secondary schools for ages 11 to 15/16 and upper secondary (lycées) to the age of 17/18. School education is compulsory from the age of six to 16 years, although

many attend pre-primary education from the age of three. There is a national curriculum for all levels set out in terms of knowledge, capacities and attitudes. As in the English National Curriculum this is expressed in terms of learning objectives. The central government determines the curriculum and provides the inspectorate, while the responsibility for costs and running of schools is at three different levels: the cities for primary schools, the departments for lower secondary schools and the regions for the lycées. Teachers are free to decide content and pedagogy, and in primary schools have more leeway in relation to the school timetable and programme than is commonly believed.

A comparison of French and English teachers of pupils in the first two years of the primary education showed some interesting contrasts between their feedback practices and how they responded to pupils at different levels of achievement (Raveaud 2004). French teachers expected all pupils to tackle the same work and their 'discourse suggested that it was better for a child's self-esteem to struggle on the same task as their classmates than to be labelled a failure by being given easier work to do' (Raveaud 2004: 206). Thus in France written judgements of on-going work were made on the same basis as summative judgements. To the teachers in England, facing pupils with tasks they cannot do would be seen as damaging to their self-esteem, but this concern is overridden in France by a desire to give all pupils the same chance. There is little written feedback of a formative nature, although in their oral remarks to pupils teachers recognise effort and prior achievement. Thus teachers' own assessment, at least on paper, is criterion referenced and summative. The work of older primary pupils will often be given marks out of 20.

All pupils are tested on entering the third year of the primary school and the first year of the lower secondary school in French and mathematics; science is to be added in 2007. The tests are provided by the Ministry of Education but are administered and marked by teachers. The purpose is diagnostic for teachers and parents, which is why they are conducted at the beginning of a year. The tests are based on the national curriculum and the outcomes are used to identify pupils' educational needs. 'Each school is responsible for conducting the analysis of its own results using the specific computer software provided and for drawing up a 'success chart' for each pupil and each form' (Bonnet 1997: 300). There is no quality assurance of the teachers' marking of the tests (Broadfoot 1994). Because of the timing, the teachers cannot be held responsible for the results and there is evidence that teachers genuinely use them to inform their teaching (Bonnet quotes Thélot (1993) in this regard). The test results influence practice by drawing attention to areas of weakness across all schools which can be addressed by teachers, as well as guiding plans for individual pupils. To give further opportunity for teachers to use assessment to inform their teaching a bank of test items covering most subjects has been made available at a variety of levels for both primary and secondary school.

Representative samples of the results from the compulsory diagnostic testing in the third year of primary and first year of lower secondary are collected and analysed centrally to provide a national picture of achievement and benchmarks for teachers. However, this is not the only data on national standards available to the Ministry. There is also a national survey of samples of pupils at the end of primary and of lower secondary school. As well as tests in all subjects, information is collected in these surveys about non-cognitive attainments, attitudes and values. Comparisons over time are made possible by including some common items from year to year.

Results for individual schools are not centrally reported; only the anonymous samples and the sampled survey findings are used centrally to report regional and national results. These results are widely distributed and used at regional level to identify those areas of the curriculum where schools may need help through professional development. At the national

level resource allocation to regions takes into account the pupils' results in order to compensate for under-achievement which may have been caused by differential socio-economic factors (Bonnet 1996).

For school evaluation, the emphasis is on self-evaluation against a set of national standard indicators. These indicators fall into four categories: input (characteristics of pupils); output indicators (pupils' achievements); resources; and school management and environment. Using the provided computer programs, schools are able to compare their profile with that of similar schools nationally or in the same region. This is considered to be a better approach to school evaluation than relying on pupils' performance in tests or examinations. School inspectors are concerned with the work of individual teachers rather the school as a whole.

Overall it is apparent that, in France, assessment is used as a tool for the improvement of education at the individual pupil level through regular testing used diagnostically, at the school level through the use of indicators, and at the national level through the central collection of information. There is an underlying belief that better assessment and evaluation and the dissemination of the information will support constructive criticism that leads to improved practice.

Themes running through the examples

Even across this limited selection of six countries outside England, there are some noteworthy themes that indicate the pros and cons of alternative assessment systems. These are considered in this section before proceeding, in section 4, to suggest how different approaches to assessment at the primary level can be evaluated and compared. A convenient structure is to consider themes relating to how the various systems provide for assessment for the main purposes of helping learning, reporting learning, monitoring achievement at regional or national levels and contributing to the evaluation of teachers and schools.

Helping learning

All of the systems either implicitly or explicitly encourage the use of assessment to help learning. Formative assessment is most explicitly built into the system in Scotland, where it has been spelled out in terms of the characteristics of teachers and schools that indicate the use of assessment to help learning (Learning and Teaching Scotland 2006). This serves to underline that formative assessment is integral to teaching and is not a matter for a formal requirement or even guidelines. All that can be done is for systems to provide opportunities for assessment to be used formatively and to avoid those features that inhibit this use. Thus several countries include the 'expectation' that information that the system makes available will indeed by used to help learning.

In France, New Zealand and Sweden, teachers are provided with information that is described as 'diagnostic', raising the question of whether 'diagnostic' is the same as 'formative'. Diagnostic information has the prime purpose of alerting teachers to the needs of pupils, but catering for those needs requires action in the form of adjusting teaching (Black and Wiliam 1998a: 2) and establishing a classroom climate where teachers and pupils together decide how to take the next steps in learning (ARG 1999: 7). So it is difficult to compare systems in terms of how well they support formative assessment, since most of them claim to do so, except in considering the impact of other elements of the system. We consider this further in section 4.

Reporting learning

In relation to internal school summative assessment, for school records and reporting to parents, the account of systems in France, Northern Ireland and Scotland make reference to

procedures for using records of achievement to plan action and not merely to report progress. This active use of records may, as in Scotland, involve pupils in self-assessment and parents in taking action. There are, however, differences across the systems in relation to the practice of giving marks or grades. This is explicitly avoided in Sweden and in Scotland when the feedback to pupils is intended to be formative.

Summative assessment is criterion-based in all of the cases considered, using criteria related to the levels of achievement identified in the curricula. In these countries, external standardised tests for individual summative assessment have not been introduced or are being phased out (as in the case of Wales and Northern Ireland). This does not mean that tests are not used; indeed it is clear that, in France, for example, pupils will experience quite frequent testing. Tests that are given to all pupils at certain points in their schooling are being used for two main purposes. In France and New Zealand, teachers are required to administer tests to all pupils at the start of certain years to provide diagnostic information about achievements in core subjects as a basis for teachers to plan appropriate learning experiences. In France, New Zealand and Scotland, teachers can use items from banks of tests and tasks to check their judgements of pupils' work for summative purposes. This acknowledges that teachers' judgements need some form of moderation since they will be assessing different work conducted in differing contexts against the level description criteria. An alternative form of moderation is group discussion of examples and the creation of exemplars of the agreed operational meaning of the criteria, as proposed in Wales. Without either of these, as in the lower years of the Swedish comprehensive school, the results of teachers' assessment are of low reliability. In these early years, however, the results do not have any 'high stakes' use.

Monitoring and evaluation

Tests are also used in France, Scotland and New Zealand for the purpose of monitoring regional and national standards, but in all cases these are quite separate from tests used by teachers as just discussed. These monitoring surveys involve a relatively small sample of pupils on each occasion and are part of an ongoing programme designed to show not only what pupils can do at any one time but also to monitor changes across the years. A large number of items can be used in a survey, with any one pupil taking only a few of them. Thus the survey can provide a good sample of the curriculum domains. Results at the class and school level are of value only when combined with others to report at regional or national levels and so cannot be used for school evaluation. These surveys are therefore described as having low stakes. In France, a sample of the tests given at the beginning of the year in the primary and first year of lower secondary school is also collected. However, as this is an anonymous sample, the results cannot be used to report on the performance of individual schools.

Wide reporting of results of the national sample surveys in France, New Zealand and Scotland enable the information to be used formatively at the system level, providing feedback that can be used to identify aspects of the curriculum that may need attention. The value of this information to schools is in focusing attention on their own practice and the performance of pupils in the areas identified as weaknesses. This use of the results encourages participation in the surveys. In this way national data are collected without adding high stakes to the assessment of pupils.

High stakes use of tests is also avoided by ensuring that the evaluation of teachers and schools for accountability is based on a range of indicators relating to the context, environment, curriculum provision and resources as well as pupil performance. In Scotland and France, and planned in Wales, such varied indicators are provided for school evaluation

and school self-evaluation. This reflects an overall aim of the systems in these and many other countries, for the assessment of pupils and evaluation of schools to provide those in schools with tools to improve their practice rather than to be used by others to control teachers and schools.

4. Evaluating assessment systems

An assessment system is made of various components which serve the main purposes and uses of assessment identified in section 1. In this section the different ways in which assessment can be carried out are discussed in terms of criteria relating to the desirable properties of the information they provide. Each component of a system needs to provide information that is valid for its purpose. Another requirement is that it should provide reliable data. Also to be taken into account is the interdependence of the various system components and the impact that assessment for one purpose may have on other assessment practices and on the curriculum and pedagogy. Further, there is the practicability of an approach, including the use of resources. Assessment can be costly, both in terms of monetary resources and the time of pupils and teachers. These four qualities can be used as evaluation criteria in seeking assessment system components that are fit for purpose.

Validity

In the context of assessment, validity refers to how well what is assessed corresponds with the processes or outcomes of learning that it is intended should be assessed. This is 'construct validity', which is generally regarded as being the overarching concept that contains within it concepts such as face, concurrent, and content validity (Messick 1989; Gipps 1994). Validity is generally considered in relation to summative assessment but it is also applicable to formative assessment.

Formative assessment

For construct validity in formative assessment the methods used should provide information about what pupils can do in relation the detailed goals of a lesson and should be interpreted by the teacher in terms of progression in the development of more general ideas or skills to that next steps can be identified. The methods the teacher uses to gather the information are likely to be a combination of direct observation, including listening to pupils' discussion, and a review of what they write or draw about their experiences.

Summative assessment

For summative assessment that is for internal school uses and is conducted by the teacher, validity will depend on the range of evidence that is used. Construct validity is likely to be greater when teachers use information from the full range of learning activities, which cover all the goals, than when the special tests or tasks are used which can only cover some of the goals of pupils' work, although such tasks and tests have a role in filling gaps in teachers' observations. The problem with relying entirely on internal tests is that teachers tend to emulate external tests in developing their own tests and their assessment practices are particularly subject to this influence when there are high stakes attached to external tests (Pollard et al 2000).

When summative assessment is for external use, that is, for grouping or selection of individual pupils or to meet requirements of national assessment policies, high validity is essential, since what is assessed contains strong messages about what is valued. When the stakes are high, however, as in England where the results of national tests are used for evaluation and accountability of teachers and schools, the requirement of high validity tends

to be compromised by the need for high reliability of the results in the interests of fairness. What this means is that what is included in the test is restricted to those learning outcomes where performance can be most easily marked as correct or incorrect. This tends to exclude outcomes that are more difficult to judge unequivocally as right or wrong, such as application of concepts, reasoning, understanding (as opposed to factual knowledge) and attitudes that are likely to influence future learning. This interaction between validity and reliability is a key point that we return to in discussion of reliability.

Accountability

Validity is also of great importance in relation to the information used for accountability. It can be argued that validity here means having information about the actions and outcomes for which teachers and schools can be held accountable. In the context of pupils' learning, teachers can be held accountable for what they do in the classroom, what learning opportunities they provide and the help they give to pupils, etc. They are not necessarily responsible for whether externally prescribed learning outcomes are achieved, since this depends on other factors over which the teacher does not have control, such as the pupils' prior learning and the many out of school influences and conditions that affect their learning. Thus teachers and schools ought to be held to account for the programme of learning opportunities that is provided and the evidence of relevant learning, but not judged solely on the level of outcomes reached by their pupils.

When rewards or sanctions are attached to results, which then acquire 'high stakes', attention is inevitably focused on maximising the outcomes that are assessed. The consequence is to focus teaching content on what is assessed, and teaching methods on transmission of this content, narrowing pupils' learning opportunities. (The education service is not the only area where practices are distorted by naïve measures of accountability; the health service in England provides many examples of this impact.) For high validity, information used in accountability should include, in addition to data on pupils' achievements, information about the curriculum and teaching methods and relevant aspects of pupils' backgrounds and of their learning histories. Various school self-evaluation guidelines provide some good examples of what this means (HMIe 2006; DfES and Ofsted 2004; Estyn 2004a and 2004b). The validity of these approaches to accountability, however, is infringed if undue weight is given to pupil performance measures.

System monitoring

For monitoring standards of pupil achievement at the regional or national levels, the most valid information describes what pupils are able to do across the full range of learning objectives in particular areas of the curriculum. The interest is not in the performance of individual pupils but in the population performance in each learning domain, such as different aspects of mathematics, or reading or other subjects. Thus validity resides in how well the domain is sampled. If the data used in monitoring is a summation of individual test results, as it is in England where national tests results are used to monitor change in national standards, then the sample of the domain is restricted to the questions that any individual pupil can answer in a test of reasonable length. This is not necessarily a good sample of the domain, and will depend quite heavily on the particular content of the test. A more valid approach is to use a far greater number of items, providing a more representative sample of the domain. Since the concern is not with the performance of individual pupils, there is no need for all pupils to be given the same items. All that is needed is for each item to be attempted by an adequate sample of the population. Sampling of this kind, where only a small proportion of pupils are selected and each only takes a sample of the full range of items, is used in international surveys (such as the OECD's PISA and the IEA surveys such

as TIMSS) and in national surveys in the Scottish Survey of Assessment (SSA) (SEED 2005b) and the Assessment of Performance Unit (APU), when this existed in England, Wales and Northern Ireland (DES/WO/DENI 1989).

Reliability

While validity refers to the kind of information used in assessment and evaluation, reliability refers to the accuracy or consistency of the information. Any observation or measurement has some error; what inaccuracy is acceptable depends on the purpose. In the context of assessment reliability is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result.

Formative assessment

In formative assessment reliability is not of concern because the evidence is both collected and used by teacher and pupil and no judgement of grade or level is involved; only the judgement of how to help a pupil take the next steps in learning. The teacher can detect and correct any mistaken judgements in on-going interaction with the pupil (Black and Wiliam 2006).

Summative assessment

In relation to summative assessment it is necessary to keep in mind the trade-off between validity and reliability mentioned earlier. Striving for high reliability can reduce validity because of preference for items and procedures that provide responses that are easily measured or judged. In the case of summative assessment for internal purposes the trade-off can be in favour of validity, since no external decisions need hang on the reported data. This would suggest that, from the arguments given above, use of teachers' judgements based on the full range of work is to be preferred to the use of tests. If the evidence is derived from regular work and is gathered over a period of time, it covers a range of opportunities for pupils to show their learning without the anxiety associated with tests. Nevertheless internal summative assessment is used to record pupil achievement and report to parents and so there needs to be some consistency in the judgements made by teachers in the same school. The approach to optimising reliability of teachers' judgements, that is considered of general benefit, is through moderation meetings where teachers discuss and apply criteria to examples of pupils' work (Good 1988; Radnor 1995; Hall and Harding 2002).

Moderation is not merely desirable but necessary when summative assessment is for use outside the school. If the results have high stakes uses, either for selection of pupils or for evaluation of teachers and schools, reliability is of the essence. Teachers' judgements are known to have low reliability when no attempt is made to provide the structure or training to assure consistency in the use of criteria (Harlen 2004). By comparison with teachers' judgements, external tests are widely considered to be more reliable and therefore to be preferred for summative assessment. However, as Wiliam (2001) and Black and Wiliam (2006) have pointed out, this assumption is not justified.

Regardless of the consistency of individual test items, the fact that a test has to be limited to a small sample of possible items means that the test as a whole is a rather poor measure for any individual pupil. This is because a different selection of items would produce a different result. Wiliam (2001) estimated the difference that this would make for the end of Key Stage tests in England. With a test of overall reliability of 0.80, this source of error would result in 32 per cent of pupils being given the wrong level. The only way to reduce this error would be to increase the length of the test, but this has only a small effect. Black and Wiliam calculate that

> if we wanted to improve the reliability of Key Stage 2 tests so that only 10 per cent of students were awarded the incorrect level, we should need to increase the length of the tests in each subject to over 30 hours.
>
> Black and Wiliam (2006), p 126

Thus the case for using tests for reporting achievement of individual pupils, based on grounds of reliability, falls apart. When we also recall that efforts to achieve high reliability of a test are at the expense of validity, then the balance of advantage falls heavily on the side of using teachers' judgements. There are several ways of raising the reliability of teachers' assessment (Harlen 1994). The examples of practice in various countries show that the most commonly used are group moderation and the use of special tests or tasks that have been tried out and calibrated as assessing certain levels of achievement for teachers to use to check their judgements. The danger of these tasks being used to replace teachers' judgements is avoided where assessment is seen as a tool for improvement and not a basis for school evaluation. Where the only purpose is to give a good account of pupils' learning outcomes, there is no incentive to inflate results or depart from intended procedures. Moreover, this use of teachers' judgement is in harmony with the practice of formative assessment, as we see in considering impact below.

Accountability

In the context of accountability, reliability refers to whether the information used is sufficiently accurate for sound and fair judgements to be made. Only part of the relevant information will be concerned with pupils' learning outcomes. For this part, the arguments above make a strong case for basing the information on moderated teachers' judgements as these provide more accurate information than external tests. For the other information that is needed, about input and process variables and resources, the evaluation carried out in the school should have checks built into the process. In some systems external checks are provided by inspectors using the same criteria.

System monitoring

The reliability of national monitoring of pupil performance depends on the reliability of individual items and on the number that are included. Using only a small number of items, designed to test individual pupils, restricts the sample of the domain that is assessed; merely collecting the same data from a larger number of pupils at the national level will not increase the reliability of the assessment of the subject domain. Less reliably assessed, but important aspects of achievement, such as application of knowledge and skills, can be monitored. These are known to be highly context-dependent (Pine et al 2006) but, because a number of such items spread across different contexts can be included in a survey, a more reliable measure can be achieved. In surveys, optimum design calls for a balance between adequate sampling of the student population and adequate sampling of the subject domain: in this perspective, blanket uniform national tests are far from optimum, being over-sampled on the population and under-sampled on the subject domain.

Impact

The word 'impact' is used here to refer to what has been identified as 'consequential validity' (Messick 1989), that is, the intended and unintended consequences of an assessment (Stobart 2006).

Formative assessment

In the case of formative assessment, the purpose is to have a positive impact on learning and indeed, as suggested earlier, the process can hardly be called formative (assessment for learning) unless this is the case. There is a growing volume of evidence, mostly from studies at the secondary level, that formative assessment does raise levels of achievement. Black et al (2003) report their own research with teachers of English, mathematics and science whose pupils achieved 'significant learning gains' following the use of assessment for learning. Black et al (2003) also cite research by Bergan et al (1991), White and Frederiksen (1998) and a review of research by Fuchs and Fuchs (1986) as providing evidence of better learning when formative assessment is built into teaching. Working with younger pupils, a positive impact of non-judgemental, 'no marks', feedback on levels of interest, effort and achievement was reported in studies by Butler (1988) and Brookhart and DeVoge (1999), while studies by Schunk (1996) have found positive impacts on achievement of self-assessment.

Summative assessment

The nature of the impact of internal summative assessment on pupils varies with its frequency as well as the range of information taken into account. In many cases grades, marks or even levels are assigned to pupils' work more often than necessary and when it would be more appropriate to provide formative feedback. Also, teachers sometimes use grades as motivation, but Brookhart and DeVoge (1999) make the point that exhorting students to work 'to get a good grade' is on the one hand motivating to pupils but on the other sets up 'a performance orientation that ultimately may decrease motivation' (p 423). Good grades are sometimes given to reward effort or good behaviour rather then only as an indication of the quality of the work. This practice amounts to using grades as rewards and punishments, as extrinsic motivation, incurring all the disadvantages for students' motivation for learning that this entails (Harlen and Deakin Crick 2003; Reay and Wiliam 1999). Internal school moderation of teachers' judgements should discourage this practice and school policies should require summative assessment only when really necessary (ARG 2006a).

As well as avoiding the practice that lead to negative impact on classroom work (as reviewed by Crooks (1988); Black and Wiliam (1998b); Harlen and Deakin Crick (2003)) action can be taken that has a positive impact. There is evidence that changing teachers' assessment can encourage a richer curriculum experience for pupils. For example, Flexer et al (1995) reported changes when teachers of third grade pupils in a school district in the USA were introduced to assessment methods using evidence from pupils' classroom performance instead of using tests. The researchers reported several effects on teachers and on pupils after a year of using these methods. Teachers were using more hands-on activities, problem solving and asking pupils for explanations. They were also trying to use more systematic observations for assessment. All agreed that the pupils had learned more and that they knew more about what their pupils knew. The teachers reported generally positive feedback from their pupils, who had better conceptual understanding, could solve problems better and explain solutions.

Such experiences underline the reality that teaching will inevitably be focused on what is assessed. When conducted by testing this impact is bound to have a narrowing effect on what is taught because, as discussed earlier, tests only sample the learning outcomes and include those outcomes more easily assessed by tests. The impact can be positive, however, as the work of Flexer et al (1995) shows, if teachers use a much wider range of assessment methods. Further evidence was provided by Hall and Harding (2002) and Hall et al (1997), who reported that the introduction of teachers' assessment in the National Curriculum

Assessment in England and Wales, was perceived by teachers as having a positive impact on pupils' learning. Their summative assessment was based on teachers' judgements across a range of pupils' work. The impact was enhanced by teachers working collaboratively towards a shared understanding of the goals and of the procedures to achieve these goals. Unfortunately the funding and opportunities for these meetings declined in the face of pressure to raise test scores and the ground that was gained (in quality of teacher assessment) in the early and mid '90s was lost (Hall and Harding 2002: 13).

Accountability

However it is the use of test results for accountability, and particularly for creating performance targets and league table of schools, that puts teachers under pressure to increase scores by teaching to the tests, giving multiple practice tests and coaching pupils in how to answer test questions rather than in using and applying their understanding more widely (Harlen and Deakin Crick 2003). Other known consequences are the de-motivation of lower achieving pupils and, for all pupils, a view of learning as product rather than process (ARG 2002b). It also leads to undue attention being focused on those who are performing just below the target level, with less attention for those who are either too far below or are already above the target level. Other evidence of impacts of testing on pupils was gathered in a survey of teachers conducted by the NUT (NUT 2006).

The additional testing proposed in England in *Making Good Progress* (DfES 2007) (see section 3), would inevitably increase the pressure on teachers and the stress on pupils. The Assessment Reform Group's response to the consultation pointed out the following:

> The status of schools will be measured by new 'progress' results as well as by their results on the existing tests. The proposal to supplement schools' income in the light or these single-level test results will further increase the pressure to give these tests priority. Thus it is clear that there is here a significant addition to existing high-stakes testing pressures. We agree that schools should be expected to aspire to improve upon the attainments of their pupils. We do not agree that this is best achieved by placing yet greater emphasis on test results.
>
> High-stakes uses of individual pupils' results are likely to distort teaching and learning. What is proposed in *Making Good Progress* is not a low-stakes 'assess when ready' model based essentially on teachers' judgements, but a high-stakes external assessment, conducted every six months in every school year, in which tests are seen as being 'underpinned' by teachers' assessment, but are nevertheless a mechanism for awarding levels without any use of such assessments. We consider that there is a grave risk that this will exacerbate the current narrowing influence that national tests have on teaching and learning… the frequency of testing will mean that the experience of pupils in every year will be dominated by these single-level tests which will be even narrower than those currently used at the end of key stages. The already considerable time spent on test-related activities (estimated at around 10% of teaching and learning time in year 6 for example) would no doubt increase.
>
> ARG response to DfES 2007

These effects are by now widely known and recognised by pupils themselves -

> Students are drilled to jump through hoops that the examiner is holding…The mechanical exam process is moulding a mechanical education.
>
> Tom Greene, a secondary school pupil, writing in The Independent, 17.8.06

- and by parents -

> For my son, and for most 10-year-olds in the country, the next nine months will be ...a sterile, narrow and meaningless exercise in drilling and cramming. It's nothing to do with the skills

> of his teacher, who seems outstanding. Nor do I blame the school. It's called preparing for Key Stage 2 SATs.
>
> Alex Benaby, writing in *The Guardian*, 10.10.06

- as well as by teachers and researchers. So it is important to ask why tests and targets based on them were introduced and what benefit they were intended to have.

System monitoring

The rationale for testing is embodied in the slogan that 'testing drives up standards'. Important evidence on this matter was collected in an extensive review by Tymms of test results in England from 1995 to 2003. Tymms (2004) made reference to data on test results from nine sources in addition to the statutory national tests for pupils at ages 11 and 13. The data from five key sources (including international surveys of achievements) were analysed to show year on year changes. The pattern that was found in national test results for eleven year olds was a rise over the first five years (1995 -1999) followed by no change from 2000 to 2003. The pattern was the same for mathematics and English. While some other data supported a rise from 1995-1999, it was noted that the data from the Trends in Mathematics and Science Surveys (TIMSS) showed no rise in mathematics over this period.

While Tymms (2004) could identify several reasons why standards of tests may have changed over this time (mainly related to how cut-off scores for levels are determined when tests change from year to year) he concluded that the effect of teaching test technique (new to pupils of this age in 1995) and of teaching to the test are very likely to have accounted for a good deal of the initial change. This conclusion is supported by trends over time in other test regimes. For example, in the USA Linn (2000) found 'a pattern of early gains followed by a levelling off' (Linn 2000:6) to be typical across states where high stakes tests are used.

The trend continues in the figures for 2006, where end of Key Stage results for English show no change from 2005 and mathematics has improved by only 1 per cent. The Government's target of 80 per cent reaching the 'required standard' has still not been reached. The results have prompted further required changes in the Government's literacy and numeracy strategies with the aim of 'driving up performance in the test in future years', while commentators have suggested that:

> a more relaxed atmosphere in schools with pupils given more time to enjoy their learning rather than being taught for the test might just be the recipe for success.
>
> Garner (2006)

However, as noted earlier, using national test results to monitor standards provides a very limited view of pupils' achievement. So we cannot really tell whether or not standards are changing. A more useful picture would be obtained by a sample survey, where teachers do not know which pupils will be tested and pupils in the same class will not in any case all be given the same items, so results would not be distorted by practising what is to be assessed. Moreover, a wide ranging survey would be able to identify areas of weakness and so facilitate better targeted remedial action.

Resources

The resources required for assessment are of two main kinds: direct costs of materials, postage, external marking, and analysis and reporting results; and indirect costs of teachers' and teaching assistants' time in preparing, giving practice and invigilating tests, and in moderating teachers' assessment. Pupils' learning time is also a key resource to be considered. In England, the direct costs of national testing are borne by the QCA, but clearly, as all other costs, these are ultimately costs to the system. The costs of summative assessment far outweigh those of implementing formative assessment.

Formative assessment

Here the main cost is in providing teachers with professional development and with good descriptions of progression in the understandings, skills and attitudes that are the goals of learning. Once in place the running costs of formative assessment are zero; time may be used differently than without formative assessment but its practice does not necessarily require more or less time overall.

Summative assessment

The resources needed for internal summative assessment are essentially those of teachers' time and pupils' learning time. When teachers' judgements are used, the process need not reduce pupils' learning time since the collection and selection of examples of work for assessment has a potential value as self-assessment. Teachers' time is needed, however, for moderation meetings, for keeping records, writing reports and talking with parents. If tests are used for external summative assessment there is a tendency for school to use tests for internal assessment and to purchase commercial tests for practice, involving direct cost to the school and taking up learning time for practice tests.

It is useful to have some idea of the scale of time used for these summative assessment activities, although any figures have to be treated with great caution. The estimation of the amount of time used for various assessment activities, for both internal and external uses, was attempted by the *Assessment Systems for the Future* project (ARG 2006a), drawing on figures from three surveys of assessment costs. These were a survey by PricewaterhouseCoopers for the QCA (QCA 2004); a survey conducted by what was then the Secondary Heads Association (SHA 2004); and one focusing on science, carried out for the Royal Society by Sheffield Hallam University (2003). Table 1 combines information for 2003 for the six years of the primary school (ARG 2006b).

Table 1. Key Stage 2: Teachers' time (in hours per year)

	Y1	Y2*	Y3	Y4	Y5	Y6
Teachers' assessment (including observation, discussion, marking)	45	53	105	105	157	157
Internal testing and preparation and use of any special tasks or commercial tests	n/a	80	96	96	96	150
National testing	n/a	20	n/a	n/a	n/a	15
Moderation	40	40	25	25	25	30
Report writing	30	30	20	20	20	20
Parents' evenings	15	15	15	15	15	15
Total	130	218	261	261	313	387

* Note that in 2003 tests were required and reported at the end of KS1

The peaks of time at the end of Key Stages are evident. But it is also clear that it is not the time spent on administering tests, but the preparation for them that is most demanding. The extra time used when external summative assessment is based on tests over that required for all other assessment activities is 100 hours in Y2, 96 for Y 3-5 and 165 for Y6. So, in Y6, 165 hours, or about 5 weeks (at 33 hours per week) would be available for teachers to use in other ways. This figure is consistent with findings of NUT (2003) research that Y6 teachers spend about 4.6 hours per week preparing for national tests.

Estimates for pupil time spent on assessment suggest that practising and taking tests occupies the equivalent of about nine days a year in Y5 and 13 in Y6 above time for all other assessment activities. Again this is time that could be used in other ways.

Accountability

Turning to resources used for accountability, it inevitably takes time to gather the kind of information that we have argued is necessary for schools to provide an account of their performance. However, when accountability is based on self–evaluation by those within the school, it serves the important function of formative evaluation that should be part of the practice of any institution. The alternative, of basing judgement of schools on the performance of pupils, leads to the negative impacts on teachers and pupils outlined earlier.

In relation to system monitoring, the economical advantage of collating achievement data already available, as in using national tests for identifying national trends, must be judged against the extent to which such data provide useful and relevant information. As we have seen, using end of Key Stage test results for this purpose is of highly questionable value. Similarly the more costly process of establishing and running surveys covering a wide range of educational outcomes has to be judged against providing more detailed feedback that can be useful not only at the policy level, but also directly to practitioners. Separating monitoring from the performance of individual students would obviate the need for central collection of student assessment data. In turn, this would set student summative assessment free from the high stakes that restrict what is taught to what is assessed, whether by tests or teachers' assessment.

5. Discussion

Even if we do not wish to go so far as to claim that what is assessed determines what is taught, it cannot be denied that, as stated at the start of this paper, it does have a large impact on pupils' education experiences. For that reason, if we are concerned to have an assessment system that supports the aims of a modern education, we need to be quite clear about what we want pupils to learn.

Whilst it is not the role of this paper to identify the curriculum objectives of primary education, it is necessary to have in mind the kinds of goals that are needed in order to prepare our pupils for their part in a rapidly changing and increasingly technological world. For this, what they learn should include (but go beyond) basic skills and knowledge. Current thinking, world-wide, emphasises the importance of helping children to develop certain skills, attitudes, knowledge and understanding, that are regarded as more important than accumulating large amounts of factual knowledge. Content knowledge can be found readily from the information sources widely available through the use of computers, and especially the internet. What are needed are the skills to access these sources and the understanding to select what is relevant and to make sense of it: pupils need understanding of broad, widely applicable concepts and the ability to use them to solve problems and make decisions in new situations. Indeed, such outcomes of education appear in statements from government departments and other organisations urging the development of citizenship, creativity and economic productivity; whilst the OECD points out that what pupils should learn in school are

> the prerequisites for successful learning in future life. These prerequisites are of both a cognitive and a motivational nature. Students must become able to organise and regulate their own learning, to learn independently and in groups, and to overcome difficulties in the learning process. This requires them to be aware of their own thinking processes and learning strategies and methods.
>
> OECD (1999) p 9

Statements such as this have implication for pedagogy, as does the emphasis on talk and interaction among pupils and between pupils and teachers (Alexander 2006).

Given that what we assess will influence whether or not pupils have opportunities to achieve such goals, it's pertinent to ask: are these important learning outcomes being assessed by the system presently in place in primary education in England? Examination of what is assessed in the national tests suggests that this is not the case. Further, since teachers' own assessment tends to follow the form and context of external assessment, this also fails to reflect these goals.

This situation is made all the more serious in the assessment system in England by using the external test results for several different purposes. What is tested for each individual pupil (mathematics and English at Key Stage 1, with science added at KS2) is also used for evaluating the performance of teachers, school and LEAs and for monitoring national standards over time. We have argued that the information from these tests is of low validity since they fail to cover some important outcomes of primary education. We have also seen that, being test-based, the information is also of low reliability, because a different selection of tests items would be likely to give different results for a significant proportion of pupils. Overall then, the current system provides information of only low dependability. Moreover, evidence of changes in standards of achievements over the years (Tymms 2004) does not support the claim that testing 'drives up standards'. Added to this is the negative impact of high stakes tests on the use of assessment to help learning, on pupils' motivation for learning, and on how the time of teachers and pupils is used. There is no indication from other countries, either, that testing improves learning; rather the reverse:

> Finland – the country with the highest standards and the smallest gap between those who do best and those who do least well – has no regular testing or inspection programme. Rather, it has a fully comprehensive, unstreamed system in which highly educated teachers… are treated as responsible professionals.
>
> Mortimer (2006)

Together these points lead to the unavoidable conclusion that the current assessment system in England is inadequate both in what is assessed and how it is being assessed. How can we do better?

The criticisms fall chiefly on the use of tests for external summative assessment and on the high stakes created by the use of results for accountability and monitoring. Given that these are separable features, since other countries make use of tests without incurring the high stakes impact, four possibilities for change are:

- Tests combined with high stakes;
- Tests with no high stakes;
- No tests, but other assessment, with high stakes;
- No tests, but other assessment, without high stakes.

Of course there are numerous other combinations, but discussion of these somewhat crude alternatives serves to highlight the principles that have to be considered.

The first of these four is what already exists in England. The second would not avoid the problem of the low validity and reliability of external tests if they are the only form of assessment to be used. The third possibility runs the risk that the alternative to tests, using teachers' judgements, could acquire the same disadvantages as the use of tests if the results were used for high stakes evaluation and monitoring. Moderation procedures would be likely to become more formalised, over-elaborate and to constrain teachers to collecting evidence using 'safe' methods that are going to 'pass' the moderation procedures. (There is evidence that this happens at secondary level where teachers' judgements are used for some parts of GCSE examinations (Donnelly et al 1993)).

So we are left with the fourth alternative, of not depending on test results and ensuring that the results of summative assessment for pupils do not acquire high stakes for the teacher and school. This would mean that information for accountability and for system monitoring would have to be provided in other ways.

The alternative to depending on test results must enable the full range of learning outcomes to be included. The use of teachers' judgements would enable this to happen since teachers can collect evidence during the numerous opportunities they have for 'observing, questioning, listening to informal discussion and reviewing written work' (ARG 2006: 9). At once this not only improves validity but removes the source of unreliability that tests cannot avoid since they can include only a narrow sample of the learning goals. A particular advantage is that teachers will be gathering this information in any case if they are using assessment for learning.

Evidence from on-going learning activities can be used both for formative assessment and for summative assessment but with an important condition – that it is reviewed and reinterpreted against the reporting criteria. The reason for this is that, as noted earlier, in formative assessment judgements are often both pupil-referenced and criterion-referenced, while for summative assessment achievement has to be criterion-references, that is, judged only against the reporting criteria. This review of the evidence needs to be done only at those times, usually twice a year, when reporting is required. Practical ways of using evidence both formatively and summatively are suggested by Harlen (2007). The point to be made here is that this approach to assessment meets the major objections to tests, assuming that effective ways of assuring quality are in place.

Several approaches to quality assurance are used in those systems that depend for summative assessment on teachers' judgements, as mentioned in section 3. The use of a calibrated bank of tasks and test items is a common one. This provides a role for tests and tasks but carries the danger that they may become the main or only source of evidence. When used to supplement teachers' judgements they have value in providing operational definitions of certain learning goals, which is of special benefit to inexperienced teachers. They can also 'plug gaps' where regular activities have, for one reason or another, not provided opportunities for teachers to judge students' performance. This is a somewhat different role than using the results of tests to give a separate assessment which is compared with that form teachers' assessment. This happens in the end of Key Stage tests in England, where

> the teachers' judgements and test results in the core subjects are reported alongside each other and are said to have equal weight. The rationale for reporting both is that they are intended to assessment different types of performance. But evidence from QCA surveys shows that many teachers include the test results in the evidence they use to form their judgements and so the value of using separate sources of evidence is compromised. In any case, teachers know that it is only the test results that matter since these are used for setting targets and evaluating schools' performance.
>
> Harlen (2007) p 144

In the Scottish system, the intention is that teachers use national tests as a means of moderating their judgements. If the results do not agree then the teacher may use evidence from the test (which he or she has administered and marked) to reconsider the decision about the level reached. But this is only one way in which teachers can moderate their judgement, and without the high stakes attached to the results in Scotland there is much less imperative to use tests.

Alternative means of quality assurance are group discussion of examples of pupils' work to align judgements, and the use of exemplars of assessed work (usually written but could

include video clips of performance). Moderation meetings, although more difficult to implement than the use of exemplars by individual teachers, have benefits beyond the reliability of the outcome:

> A system of moderation of teachers' judgements through professional collaboration benefits teaching and learning as well as assessment. Moderation that affects the planning and implementation of assessment, and consequently teachers' understanding of learning goals and of the criteria indicating progress towards them, has more than a quality assurance function.
>
> ARG (2006a) p 6

Other conditions that are known to increase the reliability of teachers' judgements are the provision of detailed criteria linked to learning goals, professional development that addresses the known sources of error and bias in teachers' judgements, and a school culture in which assessment is discussed constructively and positively and not seen as a necessary chore (Harlen 2005).

A revised system must make provision for dependable school evaluation and national monitoring. This will involve assessment of pupils since pupil performance is undeniably an essential measure of the effectiveness of an education system. However, as we have argued, there are many other influences that affect pupils' achievement and schools ought not to be judged solely on the levels of pupil performance, but on the wider range of provision they make for their pupils' education. For national monitoring, a far greater sample of performance in a domain is needed than is provided by collecting the results of individual pupils who have all taken the same test. Whilst it would be possible to collect this wider information from teachers' assessment of a sample of pupils, it would be less intrusive and more reliable for monitoring trends over time to use a regular survey. A small sample of the pupil population, between them answering a range of items, is all that is needed to provide a good estimate of pupils performance in a domain and to identify where strengths and weaknesses lie to inform policy and practice.

Finally, an effective assessment system is an open one, where all involved know what evidence is used and how it is judged. Much of the emotion aroused by assessment is a result of fear or suspicion of the unknown. To take this away we need to be completely open about the need for and purpose of assessment and why it is carried out in particular ways. Even the youngest pupils can be given some explanation of what evidence they and their teachers can use to judge the progress they are making. This helps pupils to take part in assessing their own work, which is a key feature of using assessment to help learning. It is equally important for summative assessment so that there are no surprises (for pupils or parents) in the reports of the level reached at a particular time.

References

ACCAC (Qualifications, Curriculum and Assessment Authority for Wales) (2004) *Review of the School Curriculum and Assessment Arrangements 5 – 16.* Cardiff: ACCAC.

Alexander, R.J. (2006) *Towards Dialogic Teaching: rethinking classroom talk*, 3rd edition. York: Dialogos. (First edition 2004).

ARG (Assessment Reform Group) (2006a) *The Role of Teachers in the Assessment of Learning*. Obtainable from the ARG website: www.assessment-reform-group.org and from the CPA office of the Institute of Education, University of London.

ARG (Assessment Reform Group) (2006b) ASF Working Paper 3 http://k1.ioe.ac.uk/tlrp/arg/ASF-workingpaper3.htm

ARG (Assessment Reform Group) (2002a) *Assessment for Learning: 10 Principles*. Obtainable from the ARG website: www.assessment-reform-group.org and from the CPA office of the Institute of Education, University of London.

ARG (Assessment Reform Group) (2002b) *Testing, Motivation and Learning*. Obtainable from the ARG website: www.assessment-reform-group.org and from the CPA office of the Institute of Education, University of London.

ARG (Assessment Reform Group) (1999) *Assessment for Learning: Beyond the Black Box*. Obtainable from the ARG website: www.assessment-reform-group.org and from the CPA office of the Institute of Education, University of London.

ASCL (Association for School and College Leaders) (2006) *Chartered Examiners*. Policy Paper 13. Leicester: ASCL.

Benaby, A.(2006) 'Losing a year and gaining … nothing'. *The Guardian,* October 10, 2006

Bergan, J. R., Sladeczek, I.E., Schwarz, R.D. and Smith, A.N. (1991) 'Effects of a measurement and planning system on kindergarteners' cognitive development and educational programming', *American Educational Research Journal*, 28: 683-714.

Black, P (1997) 'Whatever happened to TGAT?' in (ed) C. Cullingford, *Assessment versus Evaluation*. London: Cassell

Black, P., Harrison, C., Lee, C., Marshall, B, and Wiliam, D. (2003) *Assessment for Learning: Putting it into Practice.* Maidenhead: Open University Press.

Black, P. and Wiliam, D. (2006) 'The reliability of assessment', in J. Gardner (ed) *Assessment and Learning*. London: Sage.

Black, P. and Wiliam, D. (1998a) *Inside the Black Box.* Slough: nferNelson.

Black, P and Wiliam, D. (1998b) 'Assessment and Classroom Learning', *Assessment in Education,* 5 (1): 1-74

Bonnet, G. (1997) 'Country profile from France', *Assessment in Education,* 4 (2) 295-306

Brookhart, S. and DeVoge, J. (1999) 'Testing a theory about the role of classroom assessment in pupil motivation and achievement', *Applied Measurement in Education*, 12: 409-425.

Butler, R. (1988) 'Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance', *British Journal of Education Psychology*, 58: 1 -14.

Crooks, T.J. (1988) 'The impact of classroom evaluation practices on students', *Review of Educational Research*, 58: 438-481.

Crooks, T.J. (2002) 'Educational Assessment in New Zealand Schools', *Assessment in Education*, 9 (2) 237-254

DES/WO (1988) *Task Group on Assessment and Testing: a Report*. London: Department of Education and Science and Welsh Office.

DES/WO/DENI (1989) *National Assessment: the APU Science Approach*. London: HMSO.

DfES (2007) *Making Good Progress* Consultation. London: Department for Education and Skills

DfES (2004) *Excellence and Enjoyment: Learning and Teaching in the Primary Years.* London: DfES

DfES and OfSTED (2004) *A New Relationship with Schools: Improving Performance through School Self-Evaluation*. London: Department for Education and Skills and Office for Standards in Education.

Donnelly, J.F., Buchan, A.S., Jenkins, E.W., Welford, A.G. (1993) *Policy, Practice and Teachers' Professional Judgement: The Internal Assessment of Practical Work in GCSE Science*. Driffield: Nafferton Books

Estyn (2004a) *Guidance on the Inspection of Primary and Nursery Schools*. Cardiff: Estyn.

Estyn (2004b) *Guidance on the Inspection of Secondary Schools*. Cardiff: Estyn.

Flexer, R.J., Cumbo, K., Borko, H., Mayfield, V and Maion, S.F. (1995) *How 'messing about' with performance assessment in mathematics affects what happens in classrooms* (Technical Report 396) Los Angeles Centre for Research on Evaluation, Standards and Student Testing (CRESST).

Fuchs, L.S. and Fuchs, D. (1986) 'Effects of systematic formative evaluation: a meta-analysis', *Exceptional Children*, 53: 199-208.

Gardner, J and Cowan, P. (2005) 'The fallibility of high stakes '11 plus' testing in Northern Ireland', *Assessment in Education*, 12 (2):145-165.

Garner, R. (2006) 'Is a more relaxed atmosphere in our primary schools the key to better pupil performance?' *The Independent,* December 7 2006

Gipps, C. (1994) *Beyond Testing.* London: Falmer Press.

Good, F. J. (1988) 'Differences in marks awarded as a result of moderation: some findings from a teachers assessed oral examination in French', *Educational Review,* 40: 319 – 331.

Government of New Zealand (1998) *Assessment for Success in Primary Schools.* Wellington: Ministry of Education

Greene, T. (2006) 'There's more to education than exams', *The Independent*, Thursday 17 August 2006: 35.

Hall, K. and Harding, A. (2002) 'Level descriptions and teacher assessment in England: towards a community of assessment practice', *Educational Research,* 44: 1-15.

Hall, K., Webber, B., Varley, S, Young, V and Dorman, P. (1997) 'A study of teachers' assessment at Key Stage 1', *Cambridge Journal of Education,* 27: 107-122.

Harland, J., Moor, H., Kinder, K. and Ashworth, M. (2003) *Talking 4: The Pupil Voice on the Key Stage 4 curriculum: Report 4 of the Northern Ireland Curriculum Cohort Study.* Belfast: CCEA.

Harland, J., Moor, H., Kinder, K. and Ashworth, M. (2002) *Is the Curriculum Working? The Key Stage 3 Phase of the Northern Ireland Curriculum Cohort Study.* Slough: NFER.

Harland, J., Ashworth, M., Bower, R., Hogarth, S., Montgomery, A., Moor, H. (1999a) *Real Curriculum at the start of Key Stage 3: Report Two from the Northern Ireland Curriculum Cohort Study.* Slough: NFER.

Harland, J., Kinder, K., Ashworth, M., Montgomery, A., Moor, H. and Wilkin, A (1999b) *Real Curriculum: at the end of Key Stage 2: Report One from Northern Ireland.* Slough: NFER.

Harlen, W. (2007) *Assessment of Learning.* London: Sage.

Harlen, W. (2005) 'Trusting teachers' judgements: research evidence of the reliability and validity of teachers' assessment used for summative purposes', *Research Paper in Education,* 20 (3) 245- 270

Harlen, W. (2004) 'A systematic review of the reliability and validity of assessment by teachers used for summative purposes', in *Research Evidence in Education Library,* Issue 1, London: EPPI-Centre, Social Sciences Research Unit, Institute of Education.

Harlen, W. (1994) 'Towards quality in assessment', in (ed.) W. Harlen, *Enhancing Quality in Assessment.* London: Paul Chapman.

Harlen, W. and Deakin Crick, R. (2003) 'Testing and motivation for learning', *Assessment in Education*, 10 (2): 169-208.

Hayward, L, Kane, J and Cogan, N (2000) *Improving Assessment in Scotland: Report of the National Consultation on Assessment in Scotland.* Glasgow: University of Glasgow.

HMIe (2006) *How Good is Our School? The Journey to Excellence.* Edinburgh: HMIe. http://www.hmie.gov.uk/documents/publication/hgiosjte.pdf

Hutchinson, C. and Hayward, L. (2005) 'The journey so far: assessment for learning in Scotland', *The Curriculum Journal,* 16 (2):225 – 248.

IEA (Institute of Educational Assessors) (2006) www.ioea.org.uk.

Johnston J. and McClune, W. (2000) 'Selection project sel 5.1: Pupil motivation and attitudes - self-esteem, locus of control, learning disposition and the impact of selection on teaching and learning', in *The Effects of the Selective System of Secondary Education in Northern Ireland*: Research Papers Volume II, Bangor, Co Down: Department of Education: 1-37.

Learning and Teaching Scotland (2006) *What is an AIFL School?*

http://www.ltscotland.org.uk/Images/aifl_triagram_tcm4-232905.pdf

Leonard, M. and Davey, C. (2001) *Thoughts on the 11 plus.* Belfast: Save the Children Fund.

Linn, R.L. (2000) 'Assessments and Accountability', *Educational Researcher,* 29 (2):4-16.

Messick, S. (1989) 'Validity', in R.L. Linn (ed.) *Educational Measurement,* 3rd Edition. London: Collier Macmillan, 12-103.

Mortimer, P. (2006) 'Is "irreversible" reform really sensible?' *The Guardian*, 31 October, 2006.

NEMP (National Education Monitoring Project) (2006) See website http://nemp.otago.ac.nz/index.htm.

NUT (National Union of Teachers) (2006) *NUT Briefing: The Impact of National Curriculum Testing on Pupils*, Sept 2006.

NUT (National Union of Teachers) (2003) *The Case Against National Curriculum Tests,* September 2003.

OECD (Organisation for Economic co-operation and Development) (1999) *Measuring Student Knowledge and Skills*. Paris: OECD.

Pine, J., Aschbacher, P., Rother, E., Jones, M., McPhee. C., Martin, C., Phelps, S., Kyle, T. and Foley, B. (2006) 'Fifth graders' science inquiry abilities: a comparative study of students in hands-on and textbook curricula', *Journal of Research in Science Teaching* 43 (5): 467-484.

Pollard, A. and Triggs, P. (2000) *Policy, Practice and Pupil Experience.* London: Continuum International Publishing Group.

Pollard, A., Triggs, P., Broadfoot, P., McNess, E. and Osborn, M. (2000*) What Pupils Say: Changing Policy and Practice in Primary Education.* London: Continuum.

QCA (Qualifications and Curriculum Authority) (2004) *Financial Modelling of the English Examinations System, 2003-4*, report from PriceWaterhouseCoopers (PWC) for the QCA. QCA website.

QCA (Qualifications and Curriculum Authority) (2003) *Foundation Stage Profile Handbook*. London: QCA.

Radnor, H. A. (1996) *Evaluation of Key Stage 3 Assessment Arrangements for 1995. Final Report.* Exeter: University of Exeter.

Raveaud, M. (2004) 'Assessment in French and English infant schools: assessing the work, the child or the culture?' *Assessment in Education,* 11 (2) 193-212

Reay, D. and Wiliam, D. (1999) '"I'll be a nothing": structure, agency and the construction of identity through assessment', *British Educational Research Journal*, 25: 343-345.

Schunk, D. (1996) 'Goal and self-evaluative influences during children's cognitive skill learning', *American Educational Research Journal*, 33: 359-382.

SED (Scottish Education Department) (1991) *Assessment 5-14.* Edinburgh: SED

SEED (Scottish Executive Education Department) (2005a) *Circular 02,* June, 2005. Edinburgh: SEED.

SEED (Scottish Executive Education Department) (2005b) *Information Sheet on the Scottish Survey of Achievement.* Edinburgh: SEED.

SEED (Scottish Executive Education Department) (2004) *Assessment, Testing and Reporting 3-14: our response.* Edinburgh: SEED.

SHA (Secondary Heads Association, now the Association of School and College Leaders) (2004), http://www.ascl.org.uk/MainWebSite/Resources/Document/Policy%20paper%2013%20Chartered%20examiners%20FINAL%20priced.pdf

Sheffield Hallam University Centre for Science Education (2003) *The Cost of Assessment. A Report for the Royal Society.* Available from Centre for Science Education, Sheffield Hallam University.

Smith, E. and Gorard, S. (2005) '"They don't give us our marks": the role of formative feedback in student progress', *Assessment in Education,* 12 (1): 21- 38.

Stobart G. (2006) 'The validity of formative assessment', in J. Gardner (ed) *Assessment and Learning*. London: Sage.

Thélot, C (1993) *L'évaluation du système éducatif* . Paris: Nathan

Tymms, P. (2004) 'Are standards rising in English primary schools?' *British Educational Research Journal*, 30 (4):477-494.

White, B.Y. and Frederiksen, J.T. (1998) 'Inquiry, modeling and metacognition: making science accessible to all students', *Cognition and Instruction,* 16 (1): 3 -118.

Wikstrom, C. (2006) 'Education and assessment in Sweden', *Assessment in Education,* 13 (1) 113-128

Wiliam, D. (2001) 'Reliability, validity and all that jazz', *Education 3-13*, 29 (3): 17-21.

Wilmut, J. (2004) 'Experiences of Summative Teacher Assessment in the UK. A review conducted for the Qualifications and Curriculum Authority, unpublished ms.

APPENDIX 1

THE PRIMARY REVIEW PERSPECTIVES, THEMES AND SUB THEMES

The Primary Review's enquiries are framed by three broad perspectives, the third of which, primary education, breaks down into ten themes and 23 sub-themes. Each of the latter then generates a number of questions. The full framework of review perspectives, themes and questions is at www.primaryreview.org.uk

The Review Perspectives

P1 Children and childhood
P2 Culture, society and the global context
P3 Primary education

The Review Themes and Sub-themes

T1 Purposes and values
- T1a Values, beliefs and principles
- T1b Aims

T2 Learning and teaching
- T2a Children's development and learning
- T2b Teaching

T3 Curriculum and assessment
- T3a Curriculum
- T3b Assessment

T4 Quality and standards
- T4a Standards
- T4b Quality assurance and inspection

T5 Diversity and inclusion
- T5a Culture, gender, race, faith
- T5b Special educational needs

T6 Settings and professionals
- T6a Buildings and resources
- T6b Teacher supply, training, deployment & development
- T6c Other professionals
- T6d School organisation, management & leadership
- T6e School culture and ethos

T7 Parenting, caring and educating
- T7a Parents and carers
- T7b Home and school

T8 Beyond the school
- T8a Children's lives beyond the school
- T8b Schools and other agencies

T9 Structures and phases
- T9a Within-school structures, stages, classes & groups
- T9b System-level structures, phases & transitions

T10 Funding and governance
- T10a Funding
- T10b Governance

APPENDIX 2

THE EVIDENTIAL BASIS OF THE PRIMARY REVIEW

The Review has four evidential strands. These seek to balance opinion seeking with empirical data; non-interactive expressions of opinion with face-to-face discussion; official data with independent research; and material from England with that from other parts of the UK and from international sources. This enquiry, unlike some of its predecessors, looks outwards from primary schools to the wider society, and makes full though judicious use of international data and ideas from other countries.

Submissions

Following the convention in enquiries of this kind, submissions have been invited from all who wish to contribute. By June 2007, nearly 550 submissions had been received and more were arriving daily. The submissions range from brief single-issue expressions of opinion to substantial documents covering several or all of the themes and comprising both detailed evidence and recommendations for the future. A report on the submissions will be published in late 2007.

Soundings

This strand has two parts. The *Community Soundings* are a series of nine regionally based one to two day events, each comprising a sequence of meetings with representatives from schools and the communities they serve. The Community Soundings took place between January and March 2007, and entailed 87 witness sessions with groups of pupils, parents, governors, teachers, teaching assistants and heads, and with educational and community representatives from the areas in which the soundings took place. In all, there were over 700 witnesses. The *National Soundings* are a programme of more formal meetings with national organisations both inside and outside education. They will take place during autumn 2007 and will explore key issues arising from the full range of data thus far. They will aim to help the team to clarify matters which are particularly problematic or contested and to confirm the direction to be taken by the final report. As a subset of the National Soundings, a group of practitioners - the *Visionary and Innovative Practice (VIP) group* – is giving particular attention to the implications of the emerging evidence for the work of primary schools.

Surveys

30 surveys of published research relating to the Review's ten themes have been commissioned from 69 academic consultants in universities in Britain and other countries. The surveys relate closely to the ten Review themes and the complete list appears in Appendix 3. Taken together, they will provide the most comprehensive review of research relating to primary education yet undertaken. They will be published in thematic groups from October 2007 onwards.

Searches

With the co-operation of DfES/DCSF, QCA, Ofsted, TDA and OECD, the Review is re-assessing a range of official data bearing on the primary phase. This will provide the necessary demographic, financial and statistical background to the Review and an important resource for its later consideration of policy options.

Other meetings

In addition to the formal evidence-gathering procedures, the Review team meets members of various national bodies for the exchange of information and ideas: government and opposition representatives; officials at DfES/DCSF, QCA, Ofsted, TDA, GTC, NCSL and IRU; representatives of the teaching unions; and umbrella groups representing organisations involved in early years, primary education and teacher education. The first of three sessions with the House of Commons Education and Skills Committee took place in March 2007. Following the replacment of DfES by two separate departments, DCSF and DIUS, it is anticipated that there will be further meetings with this committee's successor.

APPENDIX 3

THE PRIMARY REVIEW INTERIM REPORTS

The interim reports, which will be released in stages from October 2007, include the 30 research surveys commissioned from external consultants together with reports on the community soundings and the submissions prepared by the Cambridge team. They are listed by Review theme below, although this will not be the order of their publication. Report titles may be subject to minor amendment.

Once published, the interim reports, together with briefings summarising their findings, may be downloaded from the Review website, www.primaryreview.org.uk .

1. *Community Soundings: report on the Primary Review regional witness sessions*
2. *Submissions received by the Primary Review*
3. *Aims and values in primary education*. Research survey 1/1 (John White)
4. *The aims of primary education: England and other countries*. Research survey 1/2 (Maha Shuayb and Sharon O'Donnell)
5. *The changing national context of primary education*. Research survey 1/3 (Stephen Machin and Sandra McNally)
6. *The changing global context of primary education*. Research survey 1/4 (Hugh Lauder, John Lowe and Dr Rita Chawla-Duggan)
7. *Children in primary schools: cognitive development*. Research survey 2/1a (Usha Goswami and Peter Bryant)
8. *Children in primary schools: social development and learning*. Research survey 2/1b (Christine Howe and Neil Mercer)
9. *Teaching in primary schools*. Research survey 2/2 (Robin Alexander and Maurice Galton)
10. *Learning and teaching in primary schools: the curriculum dimension*. Research survey 2/3 (Bob McCormick and Bob Moon)
11. *Learning and teaching in primary schools: evidence from TLRP*. Research survey 2/4 (Mary James and Andrew Pollard)
12. *Curriculum and assessment policy: England and other countries*. Research survey 3/1 (Kathy Hall and Kamil Øzerk)
13. *The impact of national reform: recent government initiatives in English primary education*. Research survey 3/2 (Dominic Wyse, Elaine McCreery and Harry Torrance)
14. *Curriculum alternatives for primary education.* Research survey 3/3 (James Conroy and Ian Menter)
15. *The quality of learning: assessment alternatives for primary education*. Research survey 3/4 (Wynne Harlen)
16. *Standards and quality in English primary schools over time: the national evidence*. Research survey 4/1 (Peter Tymms and Christine Merrell)
17. *Standards in English primary schools: the international evidence*. Research survey 4/2 (Chris Whetton, Graham Ruddock and Liz Twist).
18. *Quality assurance in primary education*. Research survey 4/1 (Peter Cunningham and Philip Raymont)
19. *Children, identity, diversity and inclusion in primary education.* Research survey 5/1 (Mel Ainscow, Alan Dyson and Jean Conteh)
20. *Children of primary school age with special needs: identification and provision*. Research survey 5/2 (Harry Daniels and Jill Porter)

21. *Children and their primary education: pupil voice*. Research survey 5/3 (Carol Robinson and Michael Fielding)

22. *Primary education: the physical environment*. Research survey 6/1 (Karl Wall, Julie Dockrell and Nick Peacey)

23. *Primary education: the professional environment*. Research survey 6/2 (Ian Stronach, Andy Pickard and Elizabeth Jones)

24. *Teachers and other professionals: training, induction and development*. Research survey 6/3 (Olwen McNamara, Rosemary Webb and Mark Brundrett)

25. *Teachers and other professionals: workforce management and reform.* Research survey 6/4 (Hilary Burgess)

26. *Parenting, caring and educating*. Research survey 7/1 (Yolande Muschamp, Felicity Wikeley, Tess Ridge and Maria Balarin)

27. *Children's lives outside school and their educational impact*. Research survey 8/1 (Berry Mayall)

28. *Primary schools and other agencies*. Research survey 8/2 (Ian Barron, Rachel Holmes, Maggie MacLure and Katherine Runswick-Cole)

29. *The structure and phasing of primary education: England and other countries*. Research survey 9/1 (Anna Eames and Caroline Sharp)

30. *Organising learning and teaching in primary schools: structure, grouping and transition*. Research survey 9/2 (Peter Blatchford, Judith Ireson, Susan Hallam, Peter Kutnick and Andrea Creech)

31. *The financing of primary education.* Research survey 10/1 (Philip Noden and Anne West)

32. *The governance, administration and control of primary education*. Research survey 10/2 (Maria Balarin and Hugh Lauder)

The Primary Review is a wide-ranging independent enquiry into the condition and future of primary education in England. It is supported by Esmée Fairbairn Foundation, based at the University of Cambridge and directed by Robin Alexander. The Review was launched in October 2006 and aims to publish its final report in autumn 2008.

FURTHER INFORMATION

www.primaryreview.org.uk

General enquiries: enquiries@primaryreview.org.uk

Media enquiries: richard@margrave.co.uk

Published by the Primary Review,
Faculty of Education, University of Cambridge
184 Hills Road, Cambridge, CB2 8PQ, UK

ISBN 978-1-906478-03-2